Transit Safety Modeling and Operational Efficiency: An Integrated Data Architecture and Forecasting Approach for Washington Metropolitan Area Transit Authority

Amna Maqsood[1], Alana Lee[1], Gladys Fong[1], Shannon Zhang[1], Amanda Redmiles[2] , Eric Dano[1,*], and Adam Jachimowicz[2]

[*] *Corresponding Author: ericdano@gwu.edu*

[1] *Department of Engineering Management and Systems Engineering*
The George Washington University
Washington D.C., USA

[2] *Department of Safety*
Washington Metropolitan Area Transit Authority
Washington D.C., USA

***Abstract*—As urban sprawl increases in major U.S. metropolitan cities, maintaining safety in large, interconnected transit systems while reducing preventable incidents is essential for central business districts like Washington D.C. Complex Adaptive Systems (CAS) like an interstate transit system are at higher risk of injuries, liability costs, and operational bottlenecks as ridership and demands grow. Safety transit operations are heavily dependent on data analytics for continuous improvement and preventable risk reduction. At the Washington Metropolitan Area Transit Authority (WMATA), many teams process and produce safety data; however, these efforts are siloed, slowing operational efficiency. The project aims to establish a Data Architecture (DA) model and forecasting tool capable of supporting advanced analytics and faster decision-making processes to increase operational efficiency. Using historical incident and safety asset data with the existing DA, we co-designed an updated architecture and forecasting tool with the data analysts' input. By standardizing the documentation of conceptual and logical data models, the links between common datasets and assets can reduce CAS data silos. The supporting forecasting tool is developed using non-linear regression, Prophet model, and analysis of geospatial correlation through Moran's I approach. Models were designed to incorporate seasonality, location, and time-based indicators found in the historical data, thereby increasing uniqueness in DA. Validation focuses on data consistency, model integration, and alignment with stakeholder safety objectives. Contributing to the broader field of transportation data management, interoperable redesign and documentation create flexible data frameworks. This enables modern transit agencies to move from reactive to predictive operational management.**

Keywords—Complex Adaptive Systems, Data Architecture, Predictive Analytics, Operational Safety Management

I. INTRODUCTION

A. Problem Statement

WMATA currently faces significant challenges in managing safety and operational data due to a fragmented landscape of disparate datasets, inconsistent data formats, and manual entry processes. Data related to vehicle monitoring, incidents, system performance, maintenance, and passenger volume are stored across multiple databases with varying standards, creating obstacles for efficient analysis and reliable reporting. This fragmented approach results in excessive time spent on data cleaning and validation, increases the risk of incorrect or misleading insights, and limits the organization's ability to perform advanced analytics or predictive safety modeling. Without a unified, structured data environment, WMATA cannot fully realize the benefits of data-driven decision-making, proactive risk mitigation, or operational excellence across its transit network.

B. Background and Literature Review

The WMATA Safety Informatics and Solutions team oversees both operational and customer safety data to support system reliability. These data are essential for identifying trends and informing safety strategies, yet they currently arrive in inconsistent formats from multiple vendors, real-time systems, and PDF reports. This fragmentation slows analysis, increases the risk of errors, and limits the usefulness of the information.

WMATA's data fragmentation challenge is not unique to the organization. It reflects a systemic issue across transit agencies. Safety data typically accumulate across operational databases. Accuracy is further compromised by manual documentation practices, such as inconsistent standards, varying collection frequencies, and disparate storage architectures [1]. The consequences of this fragmentation are measurable and span multiple systems; analysis is delayed, and the organization operates primarily in reactive mode, addressing incidents after they occur rather than anticipating them.

The transportation safety literature increasingly demonstrates that integrated, data-driven safety management is both financially and operationally justified. Safety data management systems that consolidate data across operational domains have been shown to generate quantifiable returns on investment. Cost-benefit analyses have confirmed that spending

on unified data infrastructure recovers costs through reduced liability, improved maintenance scheduling, and prevented incidents [2]. This provides the economic incentive for WMATA's integration initiative. Additionally, contemporary research shows that when integration is coupled with predictive monitoring, detection lead times improve significantly [3]. This suggests that WMATA should move past archival storage and toward active oversight.

This reorientation aligns with the Safe System Approach, a paradigm increasingly adopted in transportation safety that reframes safety as a systemic and architectural challenge rather than a collection of isolated incidents [4]. Under this framework, safety emerges from design, such as how data flows and how feedback loops function. Data and decisions flow across the organization rather than being siloed within departments.

C. Significance

The significance of this project lies in its potential to modernize WMATA's DA through the development of a unified, scalable, and adaptable architecture framework. By consolidating operational data into a single high-quality environment, the system mitigates inefficiencies and reduces the risk of inaccurate or misleading analytical outcomes caused by fragmented data sources. Standardized storage practices for improved interstation compatibility and streamlined analytical workflows will enable more reliable assessments and strengthen predictive safety initiatives. Collectively, these improvements support data-informed operational excellence, promote environmentally sustainable practices, and position WMATA to make faster, more effective decisions that enhance system-wide safety.

D. Scope and Limitations

This study develops a redesigned DA for WMATA's safety incident data, including standardized models, a canonical incident database, and a forecasting component. The work is limited to datasets available to the Safety Informatics and Solutions team and does not include full integration with WMATA's enterprise systems, real-time ingestion, or operational deployment. The data and DA we received from the Safety Informatics and Solutions team contain sensitive and proprietary information. We created numbered regional location names, incident types, and can only share the concept behind our sustainable DA. The Forecasting performance is constrained by the quality and completeness of historical data from 2022 - 2025, and broader validation across additional datasets or operational contexts remain outside the scope of this project.

II. Project Goal and Research Questions

The goal of this research is to document and develop a unified, scalable, and adaptable DA model that consolidates WMATA's incidents and accidents data into a single, high-quality environment, enabling advanced analytics, predictive modeling, and real-time decision-making across the agency's multi-modal transit network.

1) What is WMATA's current data architecture, and what architecture models are most compatible with the existing system?

2) How can predictive modeling and advanced analytics be applied to WMATA's operational data to improve safety outcomes and operational efficiency?

3) What technical and organizational challenges limit the adoption of data-driven decision-making in public transit systems?

4) What factors should be considered to build sustainable, data-driven strategies for risk reduction and resource optimization?

III. Methodology

This study employed a dual-track methodology to address WMATA's safety data challenges, encompassing a DA redesign and the development of a predictive forecasting tool. Both tracks were developed in parallel through an iterative, co-design process conducted in direct collaboration with WMATA's Safety Informatics and Solutions team and partnering contractors. Together, the two components form an integrated framework in which the redesigned architecture provides the standardized data foundation necessary to support the forecasting model's predictive capabilities.

A. Data Architecture Redesign

The redesign procedure was shaped by a structured methodology to ensure that WMATA's safety data environment could support consistent and scalable integration. First, all available datasets were validated and cleaned to remove formatting irregularities, resolve duplicated fields, and confirm alignment with existing documentation. This step was essential for establishing a reliable baseline for subsequent architectural development. Next, the team assessed end-to-end data flow across departments to identify how information was generated, transferred, and stored. This analysis was the foundation for the development of standardized naming conventions and field definitions, which were then uniformed to reduce ambiguity and improve interoperability. Using these standards, the redesigned DA was modeled in Lucidchart as an Entity-Relationship Diagram (ERD). The ERD formalized entity structures, relationships, and data movement pathways, providing a clear logical representation of the proposed architecture. The ERD model ensures that the redesigned architecture is fully aligned with the operational data needs and system workflows.

B. Forecasting Tool Development

WMATA's historical incident data required substantial pre-processing before forecasting analysis could proceed. The initial dataset contained 32.7 thousand incidents spanning 2022 through 2025, across bus, para-transit, and rail modalities. A mixed-methods approach was developed to ensure both qualitative transit alignment and statistical validity. Semi-structured interviews with the WMATA team led to the identification of Key Performance Indicators (KPIs) which are common in recurring analyses. Three main KPIs (incident type, location, and time) were integrated in the Prophet forecasting model, the geospatial correlation analysis (Moran's I), and the GIS-based interactive map tool.

The Prophet model was developed and tested in Python with all incidents aggregated at the weekly level. This aggregation

strategy offered critical methodological advantages in reduced daily volatility and stable confidence intervals for medium-range forecasts.

The geospatial correlation analysis required a minimum of 30 observations per incident type per region. Transit mode informed geospatial correlation and thus spatial weights. Bus and para-transit incidents were analyzed using a Distance Band weights matrix with a 1 km radius, where all incidents within that distance are counted as neighbors, while rail incidents were analyzed using K-Nearest Neighbors with k = 6, ensuring every station remains connected to meaningful neighbors even where stops are sparse or unevenly spaced along a fixed corridor. Significance was assessed using three measures: Moran's I, which ranges from -1 to +1 and indicates the degree of spatial clustering or dispersion; the Z-score, which measures how far the observed result falls from what would be expected under complete spatial randomness; and a permutation-based p-value derived from 999 Monte Carlo simulations. The result was considered statistically significant when the permutation p-value fell below 0.05.

To investigate fair cross-mode risk comparison, incident rates were normalized by daily ridership and revenue service hours (RSH). The standardization formula is as follows:

$$\textit{Ridership Incident Rate} = (\textit{Monthly Ridership}) / (\textit{Monthly Incident Count})) \times 10{,}000 \quad (1)$$

$$\textit{RSH Incident Rate} = (\textit{Monthly RSH}) / (\textit{Monthly Incident Count})) \times 100{,}000 \quad 2)$$

C. Map-Based Visualization Tool

The map visualization tool displays historical incident distributions across the WMATA transit network as individual points on an interactive map. This tool was developed by mapping the coordinates of incidents to existing open location data [5]. This allows filtering by year, incident type, rush hour status, season, location, and transit mode to identify patterns and support data-driven decision-making. The interactive design enables the dynamic analysis of safety incident trends, answering specific questions about where and when incidents occur most frequently across different parts of the transit system and under different operational conditions.

D. Validation of Methods

The validation procedures were focused on testing statistical rigor. The Prophet models were assessed using standard time-series diagnostics and approved by a WMATA SME. Spatial autocorrelation results were validated through permutation testing to confirm significance. The mapping interface was validated for coordinate mapping accuracy by WMATA. Together, these validation steps ensure both analytical reliability and practical usability for WMATA's safety management processes.

IV. Results

A. Data Architecture Outcomes

The analysis of WMATA's existing Safety DA revealed several structural inconsistencies that affected data quality and integration. Across the reviewed datasets, file formats were not standardized, and conversion procedures varied by source system. Field naming practices were also inconsistent, leading to repeated, ambiguous, or partially overlapping attribute labels. These factors made it difficult to trace incident records across systems and to ensure reliability to support analytical workflows. In response, the team redesigned the DA to address the disparate safety-related datasets provided by WMATA into a more coherent, centralized structure. The concept can be seen in Fig. 1. Overall, the findings indicated that the current architecture lacked a consistent framework for validation, formatting, and long-term data management.

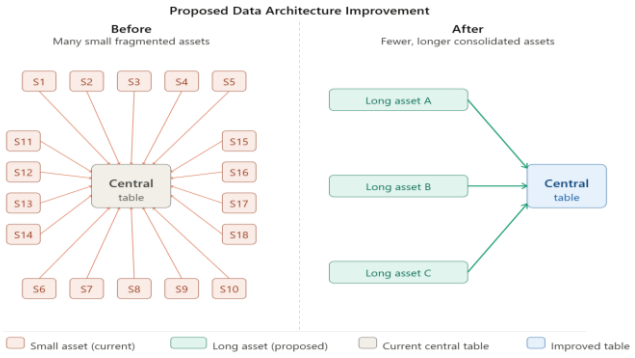

Fig. 1. Concept of Proposed Data Architecture

The redesigned model standardizes file formats and field naming conventions, aligns overlapping attributes, and defines clearer relationships between key entities: incidents, assets, locations, and corrective actions. This centralization reduces redundant data handling, improves traceability across systems, and creates a stable foundation for future reporting and analysis. The new logical DA was formalized as an Entity-Relationship Diagram (ERD) in Lucidchart. The ERD integrates the previously fragmented datasets into a single logical schema, illustrating how entities and attributes connect within the operational context. By making these relationships explicit, the model supports clearer communication and provides a practical blueprint for implementing the redesigned data environment.

B. Forecasting Model Outcomes and Performance

1) Normalization Comparison and Prophet Model

Across the 2022 to 2025 study period, bus and para-transit incident rates remained consistently and substantially higher than rail across both normalization methods, as seen in Fig. 2. Our advisors from WMATA shared how Revenue Service Hours (RSH) are often the main metric when selecting annual incident thresholds. Bus and rail began at nearly comparable RSH rates in 2022 but moved in opposite directions. The incident rates we calculated, validated our projections and revealed how safe rail transit has become despite the high ridership.

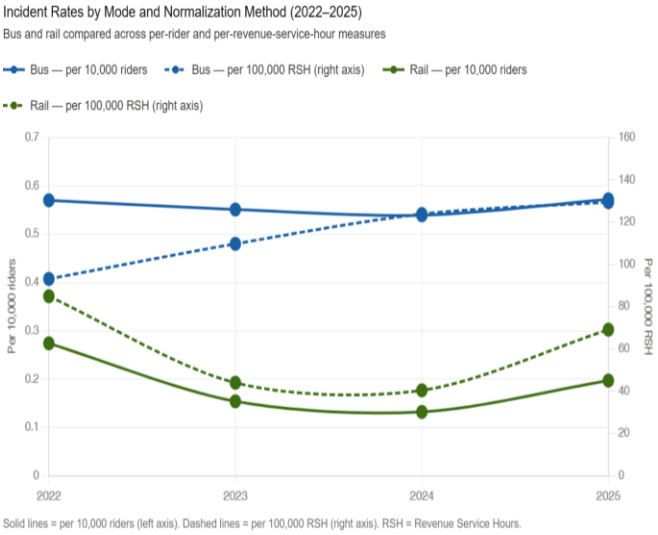

Fig. 2. Incident Rates by Mode and Normalization Comparison

Measured per 10,000 riders, bus rates remained relatively stable while rail rates declined sharply from 0.274 in 2022. The per 100,000 revenue service hours measure tells a diverging story between the two modes. Bus rates increased steadily while rail rates fell from 84.9 in 2022 to a low of 40.4.

The Prophet model decomposed the time series into trend and seasonal components, generating smoother, more interpretable forecasts optimized for executive-level visualization and operational communication, shown in Fig. 3.

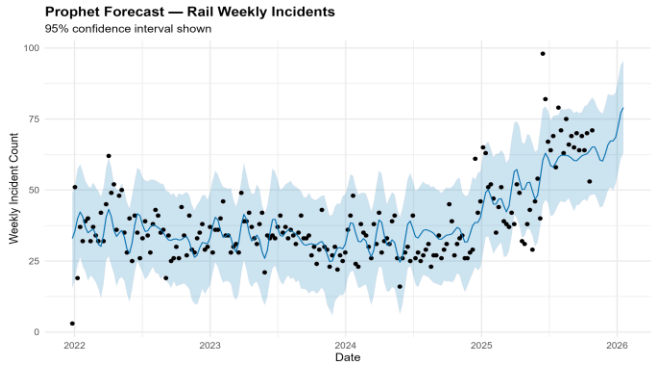

Fig. 3. WMATA Rail Prophet Forecast Results

The model displayed isolated long-term trends from recurring seasonal patterns, differentiating between genuine risk escalation and merely cyclical seasonal spikes.

2) Spatial Autocorrelation Analysis and Geographic Hotspot Detection

A critical discovery emerged when spatial autocorrelation was analyzed separately for each transit mode. Incidents do not cluster uniformly across the network; instead, they exhibit mode-specific geographic patterns requiring differentiated intervention strategies as seen in Table 1.

TABLE I. STATISTICALLY SIGNIFICANT TRANSIT MODE CLUSTERING

Transit Mode	***Geospatial Analysis Results***			
	Regions	***Moran's I***	***Z-score***	***P-perm***
Bus	*Region 1*	*0.0315*	*4.58*	*0.001*
Bus	*Region 2*	*0.0305*	*5.89*	*0.001*
Bus	*Ward A*	*0.0081*	*6.1*	*0.001*
Bus	*Ward B*	*0.0331*	*5.99*	*0.001*
Bus	*Ward C*	*0.0326*	*9.31*	*0.001*
Bus	*Ward D*	*0.0095*	*6.71*	*0.001*
Bus	*Region 3*	*0.0274*	*3.23*	*0.003*
Bus	*Ward E*	*0.0084*	*3.33*	*0.007*
Bus	*Ward F*	*0.0072*	*3.31*	*0.007*
Bus	*Region 4*	*0.031*	*1.58*	*0.048*
Bus	*Ward G*	*0.0011*	*1.17*	*0.122*
Bus	*Region 5*	*0.0018*	0.698	0.2
Bus	*Ward H*	*-0.0007*	-0.0713	*0.495*
Bus	*Region 6*	*-0.076*	-0.535	*0.676*
Bus	*Region 7*	*-0.0117*	-0.692	*0.729*

The Global Moran's I result for bus, para-transit, and rail incidents reveal a consistent pattern of statistically significant spatial clustering across much of the WMATA service area. Ten of the fifteen regions analyzed returned significant results ($p < 0.05$), with six regions including multiple DC wards and two suburban counties reaching the highest significance threshold ($p = 0.001$). Z-scores ranging from 3.23 to 9.31 across these regions confirm that the observed clustering is highly unlikely to have occurred by chance, indicating that preventable bus incidents are not randomly distributed but instead concentrate in specific geographic corridors.

The strongest clustering was observed in Ward C ($I = 0.033$, $Z = 9.31$) and Ward B ($I = 0.033$, $Z = 5.99$), suggesting these areas represent persistent hotspots for preventable bus incidents warranting targeted safety intervention. By contrast, five regions, including multiple wards and county level areas, returned non -significant results ($p > 0.05$), with two regions showing slightly negative Moran's I values, indicating a near random or mildly dispersed distribution of incidents. This geographic variation suggests that while clustering is a system-wide phenomenon for bus operations, its intensity and significance differ meaningfully across jurisdictions, pointing to localized rather than uniform risk factors across the network.

3) Incident Type and Geospatial Analysis Results

By aggregating the eight main incident types as seen in Fig. 4, we found statistically significant spatial autocorrelation ($p < 0.05$) across the WMATA service area.

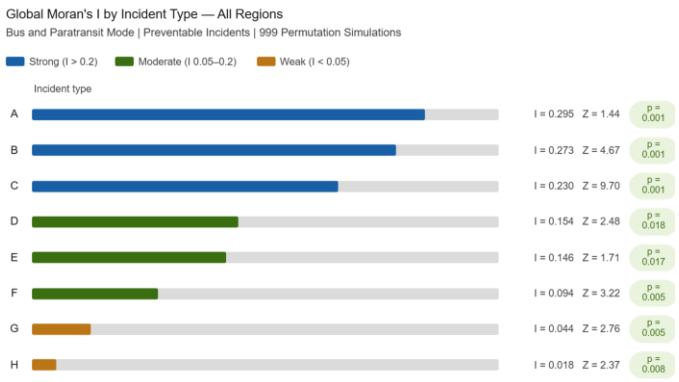

Fig. 4. WMATA Global Moran's I by Incident Type

Incident A and B both returned the strongest clustering values with permutation p-values of 0.001, followed closely by Incident C, which reached p = 0.001 and produced the highest Z-score of any type tested (Z = 9.70) across the largest regional sample. Together, these three incident types fall in the strong clustering category (I > 0.2), indicating that their geographic distribution is highly non-random at the network level.

Incidents D and E returned moderate positive spatial autocorrelation with significant permutation p-values of 0.018 and 0.017, respectively, placing them in the moderate clustering range (I = 0.05 to 0.2). Incident F, similarly, fell in the moderate range with a Z-score of 3.22 and p = 0.005, representing one of the stronger Z-scores across all types despite a lower raw I value. Incidents G and H both returned weak but statistically significant clustering (I < 0.05), with Incident H notable for having the largest sample size of any type tested (N = 2,707) while still returning a significant result at p = 0.008.

C. *Map-Based Visualization Tool Outcomes*

1.) Historical Pattern Identification

The map visualization shown in Fig. 5 identified that three specific locations consistently emerged as high-incident areas. When viewing all incidents together, we determined that 40.1% of all incidents occurred at one of three locations (Location 121, Location 2, and Location 102). This spatial concentration indicates that a disproportionate number of safety incidents occur in a limited number of locations, rather than being uniformly distributed across the network.

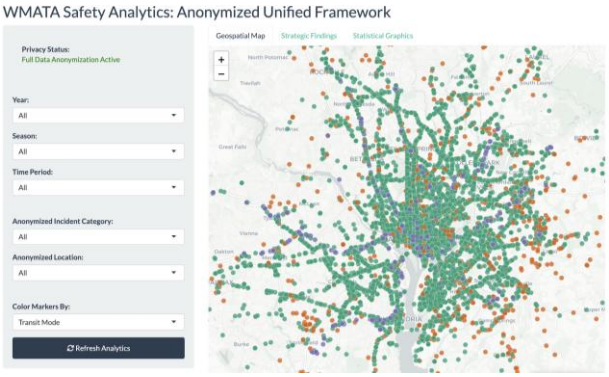

Fig. 5. WMATA Safety Dashboard, Incident Data Map Distribution

2.) Temporal Patterns and Rush Hour Dynamics

By toggling rush hour filters, the tool revealed that incident frequency varies significantly by time of day and location. For example, Location 121 shows clustering during morning rush hours, accounting for 20.7% of daily incidents in this period. This same location, however, exhibits peak clustering during evening rush hours, representing 31.8% of daily incidents during these periods. This evening peak is notably higher than morning clustering, indicating that incident risk intensifies with evening rush hours when compared to morning rush hours.

Off-peak hours show lower incident density in these same locations, accounting for 47.5% of total volume. This contrast between rush hour and off-peak incident distributions demonstrates that operational conditions during peak periods likely create elevated risk.

3.) Incident Type Classification

Filtering by incident type revealed distinct geographic patterns. For example, Incident Type 108 concentrates primarily on Location 121. This indicates location-specific risk likelihood. Looking at system-wide patterns, some incident types appear more uniformly distributed across the network, while others show similar trends to Incident Type 108. This suggests that certain incidents may be more prevalent in specific areas, while other incidents are more common across transit locations.

4.) Seasonal Variation

Filtering by season showed temporal patterns in incident risk. For example, Incident Type 108 peaks during Spring months (March–May), indicating that the frequency for this incident likely correlates with conditions occurring during the spring. Similar trends with other incident types can be observed through further exploration of the tool. Year-round, the frequency of all incidents appears to be higher in Spring and Summer months, with Fall months not being too far behind in frequency, whereas there is a drastic decrease in incidents during the Winter months, as seen in Fig. 6.

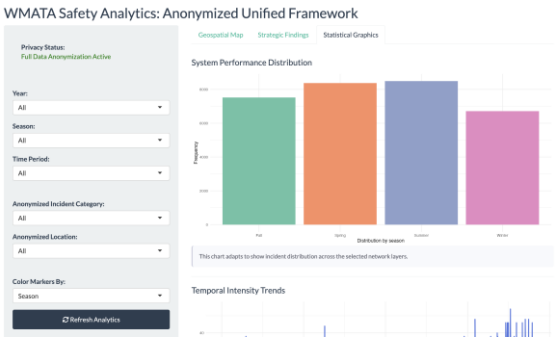

Fig. 6. WMATA Safety Dashboard, Statistical Findings

V. Conclusions and Recommendations

A. *Data Architecture*

This redesign creates a centralized, more sustainable data structure that reduces redundancy and long-term system overhead. The redesigned logical DA developed in this project addresses these challenges by consolidating the disparate safety and incident datasets into a centralized canonical incident

database and establishing standardized conceptual and logical models. The new structure aligns field names, clarifies entity relationships, and aggregates tables with overlapping or redundant fields. This centralization also reduces the need to maintain multiple disconnected data sources, supporting more efficient system operations and lowering the overall data-center footprint. Based on this work, it is recommended that WMATA adopt the canonical logical DA source for safety data and formalize agency-wide data governance standards to maintain consistency across systems. As a future enhancement, WMATA may consider implementing an automated AI-enabled normalization layer to streamline formatting, reduce manual cleaning, and further strengthen data quality. Future work should also assess how the redesigned architecture could be adapted for use by other transit agencies.

B. *Forecasting Tool*

Our forecasting models, which were based on historical trends, reveal opportunities for need-based operator training. We encourage WMATA to co-design its operator trainings with geospatial data. For example, Bus and Para-Transit operators assigned to routes in region 3 and wards H and F may benefit from targeted "refresh" trainings on incident type C, as it was the most prominent incident type for the last 4 years.

The spatial autocorrelation analysis revealed distinct patterns by transit mode: bus and para-transit incidents show statistically significant spatial clustering across the WMATA service area (Moran's I = +0.023, Z = 15.65, p = 0.001), while rail incidents show a weak positive trend that does not reach statistical significance (Moran's I = +0.084, Z = 1.44, p = 0.076), indicating that rail incidents are distributed without a detectable geographic pattern at the network level. Weekly Prophet models project upward incident trends over the medium range forecast horizon, enabling risk prediction and positioning WMATA to move from reactive incident response toward anticipatory safety management.

The normalization methods of comparison prompted us to review transit operations. We expected variation in incident rates tied to transit mode and ridership; we anticipated Rail incidents would be much higher as the Metro may have more ridership than buses. To our surprise, since late 2022, WMATA's rail has had higher average monthly ridership and lower incident rates for both normalization metrics compared to bus. Often in yearly or decade goal setting, metrics like RHS are used to set goals. We suggest that RHS, ridership, and transit modes be considered when safety goals are being assessed and set during operational planning periods.

C. *Map Tool*

The interactive mapping tool improves WMATA's ability to use safety data by making spatial and temporal patterns accessible without requiring GIS or statistical expertise. Through simple filtering by location, time, and incident type, users can quickly identify where incidents concentrate and use that information to support safety prevention planning and more efficient resource allocation. The tool also supports ongoing monitoring by allowing users to review updated incident patterns as new data are added. This capability helps WMATA detect emerging hotspots and provides a system for evaluating the impact of safety interventions over time. To strengthen this capability, it is recommended that WMATA integrate the mapping tool into routine safety reviews and continue refining location data standards. Over time, the tool can also be improved to support more advanced analytical tasks, including predictive modeling and evaluation of safety initiatives.

ACKNOWLEDGMENT

The research team thanks the Safety Informatics and Solutions team at WMATA for their guidance and access to critical data resources. We also acknowledge the Department of Engineering Management and Systems Engineering at The George Washington University for providing the tools, support, and academic framework that enabled this work.

REFERENCES

[1] A. Ungureanu, C. Doicin, C. Stanca, and A. M. Titu, "The role of an integrated quality management system in improving performance in the port organization," in Proceedings of the International Conference on Business Excellence, vol. 17, pp. 1286–1296, 2023. [Online]. Available: https://doi.org/10.2478/picbe-2023-0115. DOI: 10.2478/picbe-2023-0115

[2] P. Okwera, "A cost-benefit analysis of safety management system implementation in the transportation industry," M.S. thesis, Middle Tennessee State University, Nashville, Tennessee, 2016. [Online].

[3] D. D. Saparning, "Advancing AI-driven road safety and infrastructure resilience: A call for innovative policy and technological integration," in IISE Annual Conference. Proceedings, Norcross, pp. 1–8, 2025. [Online]. Available: https://doi.org/10.21872/2025IISE_8640. DOI: 10.21872/2025IISE_8640

[4] M. Mitman, "What matters the most for the safe system approach: Why and how to focus on kinetic energy risk," ITE Journal, vol. 94, no. 3, pp. 39–45, 2024. [Online]. Available: https://www.proquest.com/scholarly-journals/what-matters-most-safe-system-approach-why-how/docview/3039733914/se-2

[5] City of Washington, DC, "Wards from 2022," Open Data DC, 2022. [Online]. Available: https://opendata.dc.gov/api/download/v1/items/c5cd8b40fb784548a6680aead5f919ed/geojson?layers=53. [Accessed: Feb. 10, 2026].